SVKM's
D. J. Sanghvi College of Engineering

Program: B.Tech in Computer Science and Engineering (Data Science) **Academic Year: 2022** **Duration: 3 hours**
Date: 27.01.2023
Time: 09:00 am to 12:00 pm
Subject: Statistics for Data Science (Semester III) **Marks: 75**

Instructions: Candidates should read carefully the instructions printed on the question paper and on the cover page of the Answer Book, which is provided for their use.

(1) This question paper contains 7 pages.
(2) All Questions are Compulsory.
(3) All questions carry equal marks.
(4) Answer to each new question is to be started on a fresh page.
(5) Figures in the brackets on the right indicate full marks.
(6) Assume suitable data wherever required, but justify it.
(7) Draw the neat labelled diagrams, wherever necessary.

<table>
<tr><th>Question No.</th><th></th><th>Max. Marks</th></tr>
<tr><td>Q1(a)</td><td><i>Find the arithmetic mean and standard deviation by step deviation method of the following frequency distribution.</i>
<table>
<tr><td><i>Marks</i>:</td><td>0 – 20</td><td>20 – 40</td><td>40 – 60</td><td>60 – 80</td><td>80 – 100</td></tr>
<tr><td><i>Students</i>:</td><td>10</td><td>30</td><td>36</td><td>30</td><td>14</td></tr>
</table></td><td rowspan="3">[07]</td></tr>
<tr><td colspan="2">OR</td></tr>
<tr><td>Q1(a)</td><td><i>In the frequency distribution of</i> 100 <i>families given below, the number of families corresponding to expenditure groups</i> 20 – 40 <i>and</i> 60 – 80 <i>are missing from the table. However, the median is known to be</i> 50. <i>Find the missing frequencies.</i>
<table>
<tr><td><i>Expenditure</i>:</td><td>0 – 20</td><td>20 – 40</td><td>40 – 60</td><td>60 – 80</td><td>80 – 100</td></tr>
<tr><td><i>No. of families</i>:</td><td>14</td><td>?</td><td>27</td><td>?</td><td>15</td></tr>
</table></td></tr>
<tr><td>Q1(b)</td><td><i>From the following data, calculate</i> four <i>moments about assumed mean</i> 25 <i>and convert them into central moments</i>:
<table>
<tr><td><i>X</i>:</td><td>0 – 10</td><td>10 – 20</td><td>20 – 30</td><td>30 – 40</td></tr>
<tr><td><i>f</i>:</td><td>1</td><td>3</td><td>4</td><td>2</td></tr>
</table></td><td>[08]</td></tr>
<tr><td>Q2(a)</td><td><i>A distribution with unknown mean</i> μ <i>has variance equal to</i> 2.5 . <i>Use central limit theorem to find how large a sample should be taken from the dist – ribution in order that the probability will be at least</i> 0.95 <i>that the sample mean will be within</i> 0.5 <i>of the population mean.</i></td><td rowspan="3">[07]</td></tr>
<tr><td colspan="2">OR</td></tr>
<tr><td>Q2(a)</td><td><i>Among</i> 900 <i>bulbs produced</i> 90 <i>are found to be defective. Construct</i> 95 % <i>and</i> 99 % <i>confidence intervals for true proportion.</i></td></tr>
</table>

Q2(b) A genetic engineering company claims that it has developed a genetically modified tomato plant that yields on average more tomatoes than other varieties. A farmer wants to test the claim on a small scale before committing to a full − scale planting. Ten genetically modified tomato plants are grown from seeds along with ten other tomato plants. At the season's end, the resulting yields in pound are recorded as below. [08]

Sample 1 genetically modified :	20	23	27	25	25	25	27	23	24	22
Sample 2 regular :	21	21	22	18	20	20	18	25	23	20

Construct the 99% confidence interval for the difference in the population means based on these data.

Q3(a) 15.5 % of a random sample of 1600 undergraduates were smokers, whereas 20 % of a random sample of 900 postgraduates were smokers in a state. Can we conclude that less number of undergraduates are smokers than the postgraduates ? Take 5 % level of significance

OR

[07]

Q3(a) The mean breaking strength of the cables supplied by a manufacturer is 1800 with a S.D. of 100. By a new technique in the manufacturing process, it is claimed that the breaking strength of the cable has increased. To test this claim, a sample of 50 cables is tested and it is found that the mean breaking strength is 1850. Can we support the claim at 1 % level of significance ?

Q3(b) The following table shows the distribution of digits in the numbers chosen at random from a telephone directory : [08]

Digit	0	1	2	3	4	5	6	7	8	9	**Total**
Frequency	1026	1107	997	966	1075	933	1107	972	964	853	10000

Test whether the digits may be taken to occur equally frequently in the directory. Take 5 % level of significance.

Q4 In order to select suppliers of a certain type of packaging material to pack its products, a company received packaging material samples from four suppliers. Six pieces per lot are randomly selected and tested for bursting strength. The observations on the bursting strength $\left(in \frac{kg}{cm^2}\right)$ are given in the table below. Perform ANOVA at 5 % L.O.S. If hypothesis is rejected, then perform the Post − hoc Tukey's HSD test at 5 % level of significance. [15]

Observation	**Supplier**			
	A	**B**	**C**	**D**
1	12.0	10.8	11.2	12.1
2	11.5	11.0	11.5	12.5
3	12.3	11.3	12.8	12.3
4	11.9	10.7	12.1	12.0
5	12.8	10.0	11.6	11.8

6	12.1	10.7	12.0	11.8

OR

Q4 *Five treatments are used on four types of fabrics and the linear shrinkage percentage is assessed in each case. Each fabric of certain length is made into five pieces and the five treatments are randomly used. The data from this experiment are then arranged as given in the following table. It is assumed that there is no significant interaction between treatment and fabric. Perform ANOVA at* 5 % *level of significance to test whether there is any significant difference between treatments and between fabrics (blocks).*

Treatment	***Fabric***			
	1	**2**	**3**	**4**
1	17.6	19.6	18.4	19.8
2	19.2	20.4	19.8	20.7
3	17.2	19.0	17.1	17.3
4	17.0	20.1	17.1	17.7
5	17.4	18.8	17.8	16.5

Q5(a) *A simple sample of heights of* 6400 *Englishmen has a mean of* 170 *cm and a S.D. of* 6.4 *cm, while a simple sample of heights of* 1600 *Americans has a mean of* 172 *cm and a S.D. of* 6.3 *cm. Do the data indicate that Americans are, on the average, taller than the Englishmen? Take* 5 % *level of significance.*

OR

Q5(a) *Two random sample gave the following data :* [07]

Sample no.	*Size*	*Mean*	*Variance*
Sample 1	8	9.6	1.2
Sample 2	11	16.5	2.5

Can we conclude that the two samples have been drawn from the same normal population? Take 5 % *level of significance.*

Q5(b) *Ten competitors in a beauty contest were ranked by three judges as follows :* [08]

Competitors :	1	2	3	4	5	6	7	8	9	10
Ranks by Judge A $(U = u_i)$:	6	5	2	10	3	4	9	7	8	1
Ranks by Judge B $(V = v_i)$:	5	8	4	7	10	2	1	6	9	3
Ranks by Judge C $(W = w_i)$:	4	9	8	1	2	3	10	5	7	6

Discuss which pair of judges have the nearest approach to common taste of beauty.

Z - Table

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.02	0.024	0.028	0.032	0.036
0.1	0.04	0.044	0.048	0.052	0.056	0.06	0.064	0.068	0.071	0.075
0.2	0.079	0.083	0.087	0.0910	0.095	0.099	0.103	0.106	0.11	0.114
0.3	0.118	0.122	0.126	0.129	0.133	0.137	0.141	0.144	0.1480	0.152
0.4	0.155	0.159	0.163	0.166	0.1700	0.174	0.177	0.181	0.184	0.188
0.5	0.192	0.1950	0.199	0.202	0.205	0.209	0.212	0.216	0.2190	0.222
0.6	0.226	0.229	0.232	0.236	0.239	0.242	0.245	0.249	0.252	0.255
0.7	0.2580	0.261	0.264	0.267	0.27	0.273	0.276	0.279	0.282	0.285
0.8	0.288	0.2910	0.294	0.297	0.3	0.302	0.305	0.308	0.311	0.313
0.9	0.316	0.319	0.321	0.324	0.326	0.329	0.332	0.3340	0.337	0.339
1.0	0.341	0.344	0.346	0.349	0.351	0.353	0.355	0.358	0.36	0.362
1.1	0.364	0.367	0.369	0.371	0.373	0.375	0.3770	0.3790	0.3810	0.3830
1.2	0.385	0.387	0.389	0.391	0.393	0.394	0.396	0.3980	0.4	0.402
1.3	0.403	0.405	0.407	0.408	0.41	0.412	0.413	0.415	0.416	0.418
1.4	0.419	0.421	0.422	0.424	0.425	0.427	0.428	0.429	0.431	0.432
1.5	0.433	0.435	0.436	0.4370	0.438	0.439	0.441	0.442	0.443	0.444
1.6	0.445	0.446	0.447	0.448	0.45	0.451	0.452	0.453	0.454	0.455
1.7	0.455	0.456	0.457	0.458	0.459	0.46	0.461	0.462	0.463	0.463
1.8	0.464	0.465	0.466	0.466	0.467	0.468	0.469	0.469	0.47	0.471
1.9	0.471	0.472	0.473	0.473	0.474	0.474	0.4750	0.476	0.476	0.477
2.0	0.477	0.478	0.478	0.479	0.479	0.48	0.48	0.481	0.481	0.482
2.1	0.482	0.483	0.4830	0.483	0.484	0.484	0.485	0.4850	0.485	0.486
2.2	0.486	0.486	0.487	0.487	0.488	0.488	0.484	0.488	0.489	0.4890
2.3	0.489	0.49	0.49	0.49	0.49	0.491	0.491	0.491	0.491	0.492
2.4	0.492	0.4920	0.492	0.493	0.493	0.493	0.493	0.493	0.493	0.494
2.5	0.494	0.4940	0.494	0.494	0.495	0.495	0.495	0.495	0.495	0.495
2.6	0.495	0.496	0.496	0.496	0.496	0.4960	0.496	0.496	0.496	0.496
2.7	0.497	0.497	0.497	0.497	0.497	0.4970	0.497	0.497	0.497	0.497
2.8	0.497	0.498	0.498	0.498	0.498	0.498	0.498	0.498	0.4980	0.498
2.9	0.498	0.498	0.498	0.498	0.498	0.498	0.499	0.499	0.499	0.499
3.0	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.499	0.4990	0.4990

t-table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	**0.50**	**0.25**	**0.20**	**0.15**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
two-tails	**1.00**	**0.50**	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.02**	**0.01**	**0.002**	**0.001**
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

F-Table for 5 % level of significance.											
		Numerator degree of freedom									
		1	2	3	4	5	6	8	12	24	infty
Denominator degree of fredom	1	161.4	199.5	215.7	224.6	230.2	234	238.9	243.9	249	253.4
	2	18.51	19	19.16	19.25	19.3	19.33	19.37	19.41	19.45	19.5
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4	3.84	3.67
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
	8	5.32	4.46	4.07	3.87	3.69	3.58	3.44	3.28	3.12	2.93
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.9	2.71
	10	4.96	4.1	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
	11	4.84	3.98	3.59	3.36	3.2	3.09	2.95	2.79	2.61	2.4
	12	4.75	3.88	3.49	3.26	3.11	3	2.85	2.69	2.5	2.3
	13	4.67	3.8	3.41	3.18	3.02	2.92	2.77	2.6	2.42	2.21
	14	4.6	3.74	3.34	3.11	2.96	2.85	2.7	2.53	2.35	2.13
	15	4.54	3.68	3.29	3.06	2.9	2.79	2.64	2.48	2.29	2.07
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
	17	4.45	3.59	3.2	2.96	2.81	2.7	2.55	2.38	2.19	1.96
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
	19	4.38	3.52	3.13	2.9	2.74	2.63	2.48	2.31	2.11	1.88
	20	4.35	3.49	3.1	2.87	2.71	2.6	2.45	2.28	2.08	1.84
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
	22	4.3	3.44	3.05	2.82	2.66	2.55	2.4	2.23	2.03	1.78
	23	4.28	3.42	3.03	2.8	2.64	2.53	2.38	2.2	2	1.76
	24	4.26	3.4	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
	25	4.24	3.38	2.99	2.76	2.6	2.49	2.34	2.16	1.96	1.71
	26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.3	2.13	1.93	1.67
	28	4.2	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
	29	4.18	3.33	2.93	2.7	2.54	2.43	2.28	2.1	1.9	1.64
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09	1.89	1.62
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2	1.79	1.51
	60	4	3.15	2.76	2.52	2.37	2.25	2.1	1.92	1.7	1.39
	120	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83	1.61	1.25
	infty	3.84	2.99	2.6	2.37	2.21	2.09	1.94	1.75	1.52	1

$\chi^2 - table$							
Degrees of Freedom v	***Probability or L.O.S.***						
	0.99	**0.95**	**0.50**	**0.10**	**0.05**	**0.02**	**0.01**
1	0.000157	0.00393	0.455	2.706	3.841	5.214	6.635
2	0.0201	0.103	1.386	4.605	5.991	7.824	9.210
3	0.115	0.352	2.366	6.251	7.815	9.837	11.341
4	0.297	0.711	3.357	7.779	9.488	11.668	13.277
5	0.554	1.145	4.351	9.236	11.070	13.388	15.086
6	0.872	1.635	5.348	10.645	12.592	15.033	16.812
7	1.339	2.167	6.346	12.017	14.067	16.622	18.475
8	1.646	2.733	7.344	13.362	15.507	18.168	20.090
9	2.088	3.325	8.343	14.684	16.919	19.679	21.666
10	2.558	3.940	9.340	15.987	18.307	21.161	23.209
11	3.053	4.575	10.341	17.275	19.675	22.618	24.725

Q-Table for 5 % level of significance.										
		Degree of freedom for range(k)								
	df↓ k →	2	3	4	5	6	7	8	9	10
Degree of freedom for error	5	3.64	4.6	5.22	5.67	6.03	6.33	6.58	6.8	6.99
	6	3.46	4.34	4.9	5.3	5.63	5.9	6.12	6.32	6.49
	7	3.34	4.16	4.68	5.06	5.36	5.61	5.82	6	6.16
	8	3.26	4.04	4.53	4.89	5.17	5.4	5.6	5.77	5.92
	9	3.2	3.95	4.41	4.76	5.02	5.24	5.43	5.59	5.74
	10	3.15	3.88	4.33	4.65	4.91	5.12	5.3	5.46	5.6
	11	3.11	3.82	4.26	4.57	4.82	5.03	5.2	5.35	5.49
	12	3.08	3.77	4.2	4.51	4.75	4.95	5.12	5.27	5.39
	13	3.06	3.73	4.15	4.45	4.69	4.88	5.05	5.19	5.32
	14	3.03	3.7	4.11	4.41	4.64	4.83	4.99	5.13	5.25
	15	3.01	3.67	4.08	4.37	4.59	4.78	4.94	5.08	5.2
	16	3	3.65	4.05	4.33	4.56	4.74	4.9	5.03	5.15
	17	2.98	3.63	4.02	4.3	4.52	4.7	4.86	4.99	5.11
	18	2.97	3.61	4	4.28	4.49	4.67	4.82	4.96	5.07
	19	2.96	3.59	3.98	4.25	4.47	4.65	4.79	4.92	5.04
	20	2.95	3.58	3.96	4.23	4.45	4.62	4.77	4.9	5.01
	24	2.92	3.53	3.9	4.17	4.37	4.54	4.68	4.81	4.92
	30	2.89	3.49	3.85	4.1	4.3	4.46	4.6	4.72	4.82
	40	2.86	3.44	3.79	4.04	4.23	4.39	4.52	4.63	4.73
	60	2.83	3.4	3.74	3.98	4.16	4.31	4.44	4.55	4.65
	120	2.8	3.36	3.68	3.92	4.1	4.24	4.36	4.47	4.56
	infinity	2.77	3.31	3.63	3.86	4.03	4.17	4.29	4.39	4.47